Mining association rules using frequent closed itemsets

Nicolas Pasquier

Mining Association Rules using Frequent Closed Itemsets

Nicolas Pasquier
Université de Nice-Sophia Antipolis, France

INTRODUCTION

In the domain of knowledge discovery in databases and its computational part called data mining, many works addressed the problem of association rule extraction that aims at discovering relationships between sets of items (binary attributes). An example association rule fitting in the context of market basket data analysis is *cereal* $\wedge$ *milk* $\rightarrow$ *sugar* (support 10%, confidence 60%). This rule states that 60% of customers who buy cereals and sugar also buy milk, and that 10% of all customers buy all three items. When an association rule support and confidence exceed some user-defined thresholds, the rule is considered relevant to support decision making. Association rule extraction has proved useful to analyze large databases in a wide range of domains, such as marketing decision support; diagnosis and medical research support; telecommunication process improvement; Web site management and profiling; spatial, geographical, and statistical data analysis; and so forth.

The first phase of association rule extraction is the data selection from data sources and the generation of the data mining context that is a triplet D = (O, I, R), where O and I are finite sets of objects and items respectively, and R $\subseteq$ O $\times$ I is a binary relation. An item is most often an attribute value or an interval of attribute values. Each couple $(o, i) \in$ R denotes the fact that the object $o \in$ O is related to the item $i \in$ I. If an object *o* is in relation with all items of an *itemset I* (a set of items) we say that *o* contains *I*.

This phase helps to improve the extraction efficiency and enables the treatment of all kinds of data, often mixed in operational databases, with the same algorithm. Data-mining contexts are large relations that do not fit in main memory and must be stored in secondary memory. Consequently, each context scan is very time consuming.

Table 1. Example context

OID	Items
1	A C D
2	B C E
3	A B C E
4	B E
5	A B C E
6	B C E

BACKGROUND

The support of an itemset *I* is the proportion of objects containing *I* in the context. An itemset is frequent if its support is greater or equal to the minimal support threshold defined by the user. An association rule *r* is an implication with the form $r: I_1 \rightarrow I_2 - I_1$ where I_1 and I_2 are frequent itemsets such that $I_1 \subset I_2$. The confidence of *r* is the number of objects containing I_2 divided by the number of objects containing I_1. An association rule is generated if its support and confidence are at least equal to the minsupport and minconfidence thresholds. Association rules with 100% confidence are called *exact association rules*; others are called *approximate association rules*. The natural decomposition of the association rule-mining problem is:

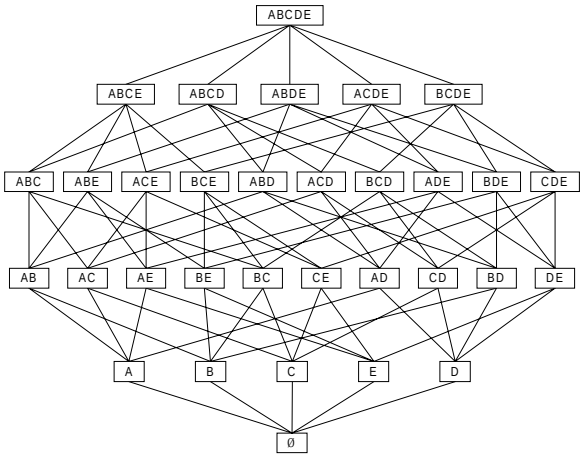
1. Extracting frequent itemsets and their support from the context.
2. Generating all valid association rules from frequent itemsets and their support.

The first phase is the most computationally expensive part of the process, since the number of potential frequent itemsets $2^{|I|}$ is exponential in the size of the set of items, and context scans are required. A trivial approach would consider all potential frequent itemsets at the same time, but this approach cannot be used for large databases where I is large. Then, the set of potential frequent itemsets that constitute a lattice called *itemset lattice* must be decomposed into several subsets considered one at a time.

Level-Wise Algorithms for Extracting Frequent Itemsets

These algorithms consider all itemsets of a given size (i.e., all itemsets of a level in the itemset lattice) at a time. They are based on the properties that all supersets of an infrequent itemset are infrequent and all subsets of a frequent itemset are frequent (Agrawal et al., 1995).

Figure 1. Itemset lattice



Using this property, the candidate *k*-itemsets (itemsets of size *k*) of the k^{th} iteration are generated by joining two frequent (*k*-1)-itemsets discovered during the preceding iteration, if their *k*-1 first items are identical. Then, one database scan is performed to count the supports of the candidates, and infrequent ones are pruned. This process is repeated until no new candidate can be generated.

This approach is used in the well known APRIORI and OCD algorithms. Both carry out a number of context scans equal to the size of the largest frequent itemsets. Several optimizations have been proposed to improve the efficiency by avoiding several context scans. The COFI* (El-Hajj & Zaïane, 2004) and FP-GROWTH (Han et al., 2004) algorithms use specific data structures for that, and the PASCAL algorithm (Bastide et al., 2000) uses a method called *pattern counting inference* to avoid counting all supports.

Algorithms for Extracting Maximal Frequent Itemsets

Maximal and minimal itemsets are defined according to the inclusion relation. Maximal frequent itemsets are frequent itemsets of which all supersets are infrequent. They form a border under which all itemsets are frequent; knowing all maximal frequent itemsets, we can deduce all frequent itemsets, but not their support. Then, the following approach for mining association rules was proposed:

1. Extracting maximal frequent itemsets and their supports from the context.
2. Deriving frequent itemsets from maximal frequent itemsets and counting their support in the context during one final scan.
3. Generating all valid association rules from frequent itemsets.

These algorithms perform an iterative search in the itemset lattice *advancing* during each iteration by one level from the bottom upwards, as in APRIORI, and by one or more levels from the top downwards. Compared to preceding algorithms, both the number of iterations and, thus, the number of context scans and the number of CPU operations carried out are reduced. The most well known algorithms based on this approach are PINCER-SEARCH (Lin & Kedem, 1998) and MAX-MINER (Bayardo, 1998).

Relevance of Extracted Association Rules

For many datasets, a huge number of association rules is extracted, even for high minsupport and minconfidence values. This problem is crucial with correlated data, for which several million association rules sometimes are extracted. Moreover, a majority of these rules bring the same information and, thus, are redundant. To illustrate this problem, nine rules extracted from the mushroom dataset (ftp://ftp.ics.uci.edu/pub/machine-learning-databases/mushroom/) are presented in the following. All have the same support (51%) and confidence (54%), and the item *free gills* in the antecedent:

1. free_gills ® edible
2. free_gills ® edible, partial_veil
3. free_gills ® edible, white_veil
4. free_gills ® edible, partial_veil, white_veil
5. free_gills, partial_veil ® edible
6. free_gills, partial_veil ® edible, white_veil
7. free_gills, white_veil ® edible
8. free_gills, white_veil ® edible, partial_veil
9. free_gills, partial_veil, white_veil ® edible

The most relevant rule from the viewpoint of the user is rule 4, since all other rules can be deduced from this one, including support and confidence. This rule is a non-redundant association rule with minimal antecedent and maximal consequent, or minimal non-redundant rule, for short.

Association Rules Reduction Methods

Several approaches for reducing the number of rules and selecting the most relevant ones have been proposed.

The application of templates (Baralis & Psaila, 1997) or Boolean operators (Bayardo, Agrawal & Gunopulos, 2000) allows selecting rules according to the user's preferences.

When taxonomies of items exist, generalized association rules (Han & Fu, 1999) (i.e., rules between

items of different levels of taxonomies) can be extracted. This produces fewer but more general associations. Other statistical measures, such as Pearson's correlation or c^2, also can be used instead of the confidence to determine the rule precision (Silverstein, Brin & Motwani, 1998).

Several methods to prune similar rules by analyzing their structures also have been proposed. This allows the extraction of rules only, with maximal antecedents among those with the same support and the same consequent (Bayardo & Agrawal, 1999), for instance.

MAIN THRUST

Algorithms for Extracting Frequent Closed Itemsets

In contrast with the (maximal) frequent itemsets-based approaches, the frequent closed itemsets approach (Pasquier et al., 1998; Zaki & Ogihara, 1998) is based on the closure operator of the Galois connection. This operator γ associates with an itemset I the maximal set of items common to all the objects containing I (i.e., the intersection of these objects). The frequent closed itemsets are frequent itemsets with $\gamma(I) = I$. An itemset C is a frequent closed itemset, if no other item $i \notin C$ is common to all objects containing C.

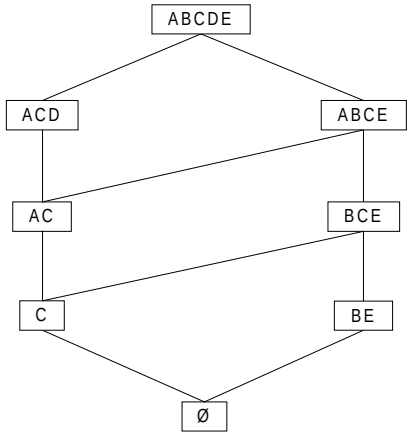
The frequent closed itemsets, together with their supports, constitute a generating set for all frequent itemsets and their supports and, thus, for all association rules, their supports, and their confidences (Pasquier et al., 1999a). This property relies on the properties that the support of a frequent itemset is equal to the support of its closure and that the maximal frequent itemsets are maximal frequent closed itemsets. Using these properties, a new approach for mining association rules was proposed:

1. Extracting frequent closed itemsets and their supports from the context.
2. Deriving frequent itemsets and their supports from frequent closed itemsets.
3. Generating all valid association rules from frequent itemsets.

The search space in the first phase is reduced to the closed itemset lattice, which is a sublattice of the itemset lattice.

The first algorithms based on this approach proposed are CLOSE (Pasquier et al., 1999a) and A-CLOSE (Pasquier et al., 1999b). To improve the extraction efficiency, both perform a level-wise search for generators of frequent closed itemsets. The generators of a closed itemset C are the minimal itemsets whose closure is C; an itemset G is a generator of C, if there is no other itemset $G' \subset G$ whose closure is C.

Figure 2. Closed itemset lattice



During an iteration k, CLOSE considers a set of candidate k-generators. One context scan is performed to compute their supports and closures; for each generator G, the intersection of all objects containing G gives its closure, and counting them gives its support. Then, infrequent generators and generators of frequent closed itemsets previously discovered are pruned. During the $(k+1)^{th}$ iteration, candidate $(k+1)$-generators are constructed by joining two frequent k-generators having identical k-1 first items.

In the A-CLOSE algorithm, generators are identified by comparing supports only, since the support of a generator is different from the supports of all its subsets. Then, one more context scan is performed at the end of the algorithm to compute closures of all frequent generators discovered.

Recently, the CHARM (Zaki & Hsiao, 2002), CLOSET+ (Wang, Han & Pei, 2003) and BIDE (Wang & Han, 2004) algorithms have been proposed. These algorithms efficiently extract frequent closed itemsets but not their generators. The TITANIC algorithm (Stumme et al., 2002) can extract frequent closed sets according to different closures, such as functional dependencies or Galois closures, for instance.

Comparing Execution Times

Experiments conducted on both synthetic and operational datasets showed that (maximal) frequent itemsets-based approaches are more efficient than closed itemsets-based approaches on weakly correlated data, such as market-basket data. In such data, nearly all frequent itemsets also are frequent closed itemsets (i.e., closed itemset lattice and itemset lattice are nearly identical),

and closure computations add execution times.

Correlated data constitute a challenge for efficiently extracting association rules, since the number of frequent itemsets is most often very important, even for high minsupport values. On these data, few frequent itemsets are also frequent closed itemsets. Thus, the closure helps to reduce the search space; fewer itemsets are tested, and the number of context scans is reduced. On such data, maximal frequent itemsets-based approaches suffer from the time needed to compute frequent itemset supports that require accessing the dataset. With the closure, these supports are derived from the supports of frequent closed itemsets without accessing the dataset.

Extracting Bases for Association Rules

Bases are minimal sets, with respect to some criteria, from which all rules can be deduced with support and confidence. The Duquenne-Guigues and the Luxenburger basis for global and partial implications were adapted to association rule framework in Pasquier et al. (1999c) and Zaki (2000). These bases are minimal regarding the number of rules; no smaller set allows the deduction of all rules with support and confidence. However, they do not contain the minimal non-redundant rules.

An association rule is redundant, if it brings the same information or less general information than those conveyed by another rule with identical support and confidence. Then, an association rule r is a minimal non-redundant association rule, if there is no association rule r' with the same support and confidence whose antecedent is a subset of the antecedent of r and whose consequent is a superset of the consequent of r. An inference system based on this definition was proposed in Cristofor and Simovici (2002).

The Min-Max basis for exact association rules contains all rules $G \rightarrow g(G) - G$ between a generator G and its closure $\gamma(G)$ such that $\gamma(G) \neq G$. The Min-Max basis for approximate association rules contains all rules $G \rightarrow C - G$ between a generator itemset G and a frequent closed itemset C that is a superset of its closure: $\gamma(G) \subset C$. These bases, also called informative bases, contain, respectively, the minimal non-redundant exact and approximate association rules. Their union constitutes a basis for all association rules: They all can be deduced with their support and confidence (Bastide et al., 2000). The objective is to capture the essential knowledge in a minimal number of rules without information loss.

Algorithms for determining generators, frequent closed itemsets, and the min-max bases from frequent itemsets and their supports are presented in Pasquier et al. (2004).

Comparing Sizes of Association Rule Sets

Results of experiments conducted on both synthetic and operational datasets show that the generation of the bases can reduce substantially the number of rules.

For weakly correlated data, very few exact rules are extracted, and the reduction for approximate rules is in the order of five for both the min-max and the Luxenburger bases.

For correlated data, the Duquenne-Guigues basis reduces exact rules to a few tens; for the min-max exact basis, the reduction factor is about some tens. For approximate association rules, both the Luxenburger and the min-max bases reduce the number of rules by a factor of some hundreds.

If the number of rules can be reduced from several million to a few hundred or a few thousand, visualization tools such as templates and/or generalization tools such as taxonomies are required to explore so many rules.

FUTURE TRENDS

Most recent researches on association rules extraction concern applications to natural phenomena modeling, gene expression analysis (Creighton & Hanash, 2003), biomedical engineering (Gao, Cong et al., 2003), and geospatial, telecommunications, Web and semi-structured data analysis (Han et al., 2002). These applications most often require extending existing methods. For instance, to extract only rules with low support and high confidence in semi-structured (Cohen et al., 2001) or medical data (Ordonez et al., 2001), to extract temporal association rules in Web data (Yang & Parthasarathy, 2002) or adaptive sequential association rules in long-term medical observation data (Brisson et al., 2004). Frequent closed itemsets extraction also is applied as a conceptual analysis technique to explore biological (Pfaltz & Taylor, 2002) and medical data (Cremilleux, Soulet & Rioult, 2003).

These domains are promising fields of application for association rules and frequent closed itemsets-based techniques, particularly in combination with other data mining techniques, such as clustering and classification.

CONCLUSION

Next-generation data-mining systems should answer the analysts' requirements for high-level ready-to-use knowl-

edge that will be easier to exploit. This implies the integration of data-mining techniques in DBMS and domain-specific applications (Ansari et al., 2001). This integration should incorporate the use of knowledge visualization and exploration techniques, knowledge consolidation by cross-analysis of results of different techniques, and the incorporation of background knowledge, such as taxonomies or gene annotations for gene expression data, for example, in the process.

REFERENCES

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., & Verkamo, A.I. (1995). Fast discovery of association rules. Advances in knowledge discovery and data mining. AAAI/MIT Press.

Ansari, S., Kohavi, R., Mason, L., & Zheng, Z. (2001). Integrating e-commerce and data mining: Architecture and challenges. *Proceedings of the ICDM Conference*.

Baralis, E., & Psaila, G. (1997). Designing templates for mining association rules. *Journal of Intelligent Information Systems*, *9*(1), 7-32.

Bastide, Y., Pasquier, N., Taouil, R., Lakhal, L., & Stumme, G. (2000). Mining minimal non-redundant association rules using frequent closed itemsets. *Proceedings of the DOOD Conference*.

Bastide, Y., Taouil, R., Pasquier, N., Stumme, G., & Lakhal, L. (2000). Mining frequent closed itemsets with counting inference. *SIGKDD Explorations*, *2*(2), 66-75.

Bayardo, R.J. (1998). Efficiently mining long patterns from databases. *Proceedings of the SIGMOD Conference*.

Bayardo, R.J., & Agrawal, R. (1999). Mining the most interesting rules. *Proceedings of the KDD Conference*.

Bayardo, R.J., Agrawal, R., & Gunopulos, D. (2000). Constraint-based rule mining in large, dense databases. *Data Mining and Knowledge Discovery*, *4*(2/3), 217-240.

Brisson, L., Pasquier, N., Hebert, C., & Collard, M. (2004). HASAR: Mining sequential association rules for atherosclerosis risk factor analysis. *Proceedings of the PKDD Discovery Challenge*.

Cohen, E., et al. (2001). Finding interesting associations without support pruning. *IEEE Transaction on Knowledge and Data Engineering*, *13*(1), 64,78.

Creighton, C., & Hanash, S. (2003). Mining gene expression databases for association rules. *Bioinformatics*, *19*(1), 79-86.

Cremilleux, B., Soulet, A., & Rioult, F. (2003). Mining the strongest emerging patterns characterizing patients affected by diseases due to atherosclerosis. *Proceedings of the PKDD Discovery Challenge*.

Cristofor, L., & Simovici, D.A. (2002). Generating an informative cover for association rules. *Proceedings of the ICDM Conference*.

El-Hajj, M., & Zaïane, O.R. (2004). COFI approach for mining frequent itemsets revisited. *Proceedings of the SIGMOD/DMKD Workshop*.

Gao Cong, F.P., Tung, A., Yang, J., & Zaki, M.J. (2003). CARPENTER: Finding closed patterns in long biological datasets. *Proceedings of the KDD Conference*.

Han, J., & Fu, Y. (1999). Mining multiple-level association rules in large databases. *IEEE Transactions on Knowledge and Data Engineering*, *11*(5), 798-804.

Han, J., Pei, J., Yin, Y., & Mao, R. (2004). Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data Mining and Knowledge Discovery*, *8*(1), 53-87.

Han, J., Russ, B., Kumar, V., Mannila, H., & Pregibon, D. (2002). Emerging scientific applications in data mining. *Communications of the ACM*, *45*(8), 54-58.

Lin, D., & Kedem, Z.M. (1998). PINCER-SEARCH: A new algorithm for discovering the maximum frequent set. *Proceedings of the EBDT Conference*.

Ordonez, C., et al. (2001). Mining constrained association rules to predict heart disease. *Proceedings of the ICDM Conference*.

Pasquier, N., Bastide, Y., Taouil, R., & Lakhal, L. (1998). Pruning closed itemset lattices for association rules. *Proceedings of the BDA Conference*.

Pasquier, N., Bastide, Y., Taouil, R., & Lakhal, L. (1999a). Efficient mining of association rules using closed itemset lattices. *Information Systems*, *24*(1), 25-46.

Pasquier N., Bastide, Y., Taouil, R., & Lakhal, L. (1999b). Discovering frequent closed itemsets for association rules. *Proceedings of the ICDT Conference*.

Pasquier, N., Bastide, Y., Taouil, R., & Lakhal, L. (1999c). Closed set based discovery of small covers for association rules. *Proceedings of the BDA Conference*.

Pasquier, N., Taouil, R., Bastide, Y., Stumme, G., & Lakhal, L. (2004). Generating a condensed representation for association rules. *Journal of Intelligent Information Systems*.

Pfaltz J., & Taylor C. (2002, July). Closed set mining of biological data. *Proceedings of the KDD/BioKDD Conference*.

Silverstein, C., Brin, S., & Motwani, R. (1998). Beyond market baskets: Generalizing association rules to dependence rules. *Data Mining and Knowledge Discovery, 2*(1), 39-68.

Stumme, G., Taouil, R., Bastide, Y., Pasquier, N., & Lakhal, L. (2002). Computing iceberg concept lattices with TITANIC. *Data and Knowledge Engineering, 42*(2), 189-222.

Wang, J., & Han, J. (2004). BIDE: Efficient mining of frequent closed sequences**.** *Proceedings of the ICDE Conference*.

Wang, J., Han, J., & Pei, J. (2003). CLOSET+: Searching for the best strategies for mining frequent closed itemsets. *Proceedings of the KDD Conference*.

Yang, H., & Parthasarathy, S. (2002). On the use of constrained associations for Web log mining. *Proceedings of the KDD/WebKDD Conference*.

Zaki, M.J. (2000). Generating non-redundant association rules. *Proceedings of the KDD Conference*.

Zaki, M.J., & Hsiao, C.-J. (2002). CHARM: An efficient algorithm for closed itemset mining. *Proceedings of the SIAM International Conference on Data Mining*.

Zaki, M.J., & Ogihara, M. (1998). Theoretical foundations of association rules. *Proceedings of the SIGMOD/DMKD Workshop*.

KEY TERMS

Association Rules: An implication rule between two itemsets with statistical measures of range (support) and precision (confidence).

Basis for Association Rules: A set of association rules that is minimal with respect to some criteria and from which all association rules can be deduced with support and confidence.

Closed Itemset: An itemset that is a maximal set of items common to a set of objects. An itemset is closed if it is equal to the intersection of all objects containing it.

Frequent Itemset: An itemset contained in a number of objects at least equal to some user-defined threshold.

Itemset: A set of binary attributes, each corresponding to an attribute value or an interval of attribute values.